Soft Comput (2016) 20:2967–2979
DOI 10.1007/s00500-015-1925-9

FOCUS

Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology

Pritesh Mistry[1] · **Daniel Neagu**[1] · **Paul R. Trundle**[1] · **Jonathan D. Vessey**[2]

Published online: 20 November 2015

Abstract Drug vehicles are chemical carriers that provide beneficial aid to the drugs they bear. Taking advantage of their favourable properties can potentially allow the safer use of drugs that are considered highly toxic. A means for vehicle selection without experimental trial would therefore be of benefit in saving time and money for the industry. Although machine learning is increasingly used in predictive toxicology, to our knowledge there is no reported work in using machine learning techniques to model drug-vehicle relationships for vehicle selection to minimise toxicity. In this paper we demonstrate the use of data mining and machine learning techniques to process, extract and build models based on classifiers (decision trees and random forests) that allow us to predict which vehicle would be most suited to reduce a drug's toxicity. Using data acquired from the National Institute of Health's (NIH) Developmental Therapeutics Program (DTP) we propose a methodology using an area under a curve (AUC) approach that allows us to distinguish which vehicle provides the best toxicity profile for a drug and build classification models based on this knowledge. Our results show that we can achieve prediction accuracies of 80 % using random forest models whilst the decision tree models produce accuracies in the 70 % region. We consider our methodology widely applicable within the scientific domain and beyond for comprehensively building classification models for the comparison of functional relationships between two variables.

Keywords Big data in toxicology · Classification · Vehicle-toxicity modelling · Area under the curve · Decision tree · Random forest · Data mining

Communicated by D. Neagu.

Electronic supplementary material The online version of this article (doi:10.1007/s00500-015-1925-9) contains supplementary material, which is available to authorized users.

✉ Pritesh Mistry
p.mistry6@bradford.ac.uk

[1] Artificial Intelligence Research Group, Faculty of Engineering and Informatics, University of Bradford, Bradford BD7 1DP, UK

[2] Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Holbeck, Leeds LS11 5PS, UK

1 Introduction

Pharmaceutical drug toxicity is a major concern facing the drug discovery industry today, resulting in not only high drug attrition rates but increased developmental costs (Basavaraj and Betageri 2014), which may not be recovered if the drug does not make it to market. Of particular concern are the anticancer compounds for which it is estimated that 95 % of the attrition rates are related to toxicity concerns (Hutchinson and Kirk 2011), which is unsurprising given the nature of their modes of action as cytotoxics. Safe, effective use of anticancer drugs is often limited by their dose-dependent toxicities typically associating such drugs with a narrow therapeutic index (NTI) (Liang et al. 2013). In a study involving hospitalised patients using NTI drugs compared to non NTI drugs, it was shown that more drug-related problems (DRPs) were associated with NTI drugs, of which adverse drug reactions was one of the eight categories considered to be a DRP (Blix et al. 2010).

Toxicity prediction without experimental effort is a major area of study in the field of toxicology. The ability to predict the nature of the toxicity of a previously untested chemical structure is highly desirable to a drug discovery team deciding which chemical structures they should continue developing. The ability to make informed decisions can not only save time

and money but also bring new drugs to market sooner, benefiting the patient population as a whole. Using historical in vivo and in vitro experimental data for in silico predictions is a challenging problem. Making such predictions involves soft computing approaches that use quantitative structure activity relationships (QSAR's) to relate the activity (toxicity) of a chemical to its structure (Eldred and Jurs 1999; Eldred et al. 1999). Many useful computational models have been developed with varying success. Models ranging from studies on aquatic toxicity that utilise multiple linear regression and neural networks (Eldred et al. 1999) through to predictive hepatotoxicity (Cruz-Monteagudo et al. 2008; Low et al. 2011 and mutagenicity (Bakhtyari et al. 2013; Xu et al. 2012; Modi et al. 2012) that use k-nearest neighbour, support vector machines and random forests are frequently reported in literature.

In this study we report on a methodology developed to build QSAR classification models based on drug-vehicle toxicity relationships. As more data becomes publicly available new methodologies are required to extract the relevant information that enables successful predictive models to be built. Using data frequently reported in toxicity studies we utilise existing data and develop a methodology that is applicable across the wider domain which could be adapted for other similar predictive problems. Specifically we employ an area under the curve (AUC) approach to discriminate between the toxicity classes of different drug vehicles. A binary classification model is built using C4.5 decision tree (DT) and random forest (RF) algorithms as machine learning methods with predictive accuracies in the 70 and 80 % range respectively. That is, our models are able to predict which vehicle will provide the better toxicity relief to a particular chemical drug.

The remainder of this paper is arranged as follows: Sect. 2 provides some background information that briefly describes the purpose of drug vehicles and how they are able to reduce a drug's toxicity. Section 3 describes our proposed approach to building classifiers based on our AUC methodology. Section 4 discusses the dataset we acquired, its content and the challenges faced when using it. Section 5 provides the experimental details of our work. Section 6 reports the results generated from our predictive models and provides the discussion to this work. Section 7 ends this study with our conclusion.

2 Background

To overcome problems relating to the toxicity of a drug significant efforts are made. Some involve the use of drug vehicle formulations that can considerably attenuate this toxicity, allowing the safer use of a potent yet highly toxic compound. Drug vehicles behave as "carriers" that aid the administration and distribution of a drug through an organism and with considered selection can maximise the efficacy and minimise the side effects. For instance, studies have shown that vehicles can be selected to improve the solubility and/or permeability of a drug or to target the drug to a specific disease site (Loftsson 1998; Porter et al. 2007; Shin et al. 2009; Hans and Lowman 2002; Liu et al. 2011; Lee et al. 2003). Vehicles have also been designed and formulated to mitigate unwanted toxicities associated with drugs that are considered highly toxic (Huo et al. 2010; Uchino et al. 2005; Kelava et al. 2010). Vehicles offer toxicity relief in several ways. A vehicle that improves the solubility or permeability of a drug removes this rate limiting step (Savjani et al. 2012) in the drug's absorption profile and so may provide some form of local toxicity relief, that is, toxicity at a local site, typically the site of administration. For instance, studies on the antifungal drug amphotericin B have shown that its precipitation upon infusion is related to its acute toxicity. Solubilisation of amphotericin B using polymeric micelles, has resulted in a reduced haemolytic activity of the drug (Yu et al. 1998). Similarly the use of cyclodextrins to enhance the gastrointestinal absorption of tacrolimus was shown to reduce renal and neural toxicity in rats (Arima et al. 2001).

A drug that is toxic to a particular organ may be formulated to avoid accumulation within that organ, or targeted to a specific disease site such as a tumour thus avoiding accumulation within other organs. This type of vehicle formulation involves the modulation of a drug's plasma concentration, ensuring it remains below the toxicity threshold but above a level needed for efficacy. Vehicles that alter a drug's pharmacokinetic properties can reduce toxicity by limiting the $C_{\max}$ (peak plasma concentration) (Kim et al. 2001; Italia et al. 2007). Two drug vehicles may exhibit the same area under the plasma-drug concentration curve for the same drug, but the vehicle associated with a higher $C_{\max}$ value may produce a toxic outcome compared to a vehicle which produces a flatter and more prolonged plasma drug concentration profile (Fig. 1).

For a drug to produce observable toxicity effects it must exceed a toxicity threshold. The use of vehicle formulations can significantly alter the pharmacokinetic profile of a drug to prevent it exceeding this toxicity threshold. Figure 1 conceptualises this using the hypothetical drug X. While vehicle A produces a spike in the plasma-drug concentration curve and exceeds the toxicity threshold, vehicle B produces a flatter, prolonged plasma-drug concentration curve which remains below the toxicity threshold. Vehicle B therefore does not result in any measurable toxicity outcome but can still provide therapeutic benefit.

An alternative method for modulating a drug's toxicity is with the aid of co-administrative agents. Such agents can be other drugs, protein extracts, vitamins or oils. They work not by altering a drug's pharmacokinetic profile but instead by countering or moderating the drug's toxic manifestation

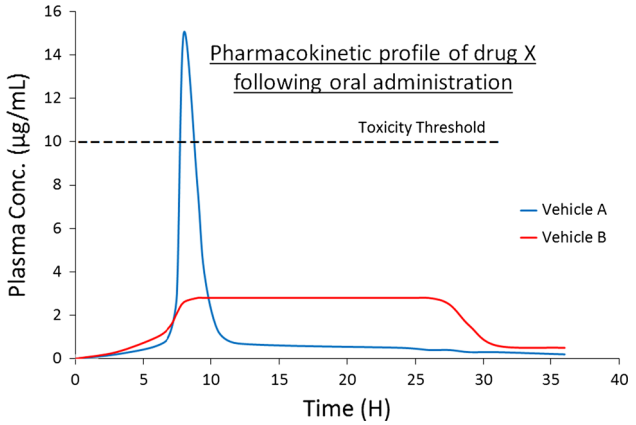

Fig. 1 Area under the plasma-drug concentration curve for drug X formulated using vehicles A and B. Vehicle A results in a concentration spike at approximately 8 h which exceeds the toxicity threshold, whilst the same drug formulated using vehicle B produces a flatter more prolonged course of absorption that does not exceed the toxicity threshold but still provides an equivalent if not better total exposure compared to vehicle A

and providing some form of toxicity relief. For example this relief may come in the form of antioxidant benefits related to these co-administrative agents. Vitamins C and E are known to provide nephroprotective effects to cisplatin induced kidney injury in mice (Ajith et al. 2007; Maliakel et al. 2008). Other co-administrative agents may behave as deactivators of toxicity pathways that are initiated upon drug administration. Acetaminophen is a globally used analgesic that results in hepatic injury at high doses. It is not the parent drug but rather its reactive metabolite *N*-acetyl-*p*-benzoquinone imine (NAPQI) that results in hepatic injury (James et al. 2003; Hodgman and Garrard 2012). Whilst at low acetaminophen doses, NAPQI is deactivated via conjugation to endogenous glutathione, at high doses glutathione becomes depleted leaving the reactive metabolite NAPQI to bind covalently to hepatic proteins resulting in hepatotoxicity. In humans the most commonly associated cytochrome P450 enzymes that result in NAPQI production are CYP2E1, CYP1A2 and CYP3A4, of which CYP2E1 accounts for the greatest transformation (Lee et al. 1996; Patten et al. 1993). Lee et al. investigated the co-administrative use of chlormethiazol, an inhibitor of CYP2E1, with high doses of acetaminophen. They demonstrated that chlormethiazol was able to prevent liver injury in mice resulting in the survival of mice administered a high dose (500 mg/kg) of acetaminophen. A greater than 50 % death rate was observed in mice that were not co-administered chlormethiazol (Lee et al. 1999).

2.1 Measuring toxicity

Experimental toxicity is measured in several ways depending on the chosen outcome of the study. For acute toxicity the lethal dose 50 (LD50) or maximum tolerated dose (MTD) is used (Kim et al. 2001; Kaminskas et al. 2012; Larabi et al. 2004). Whilst for specific organ toxicity, irregularities in organ function such as the glomerular filtration rate (GFR) for the kidneys (Mora et al. 2003) or the QT interval for the heart can be measured (Pereverzeva et al. 2007). Organ weights or total animal body weights compared to a control can also be useful indications of toxicity (Pereverzeva et al. 2007; Injac et al. 2008).

Organ specific biomarkers are frequently reported when studying toxicity relating to a specific organ. Using biomarkers is advantageous because it allows toxicity measurements to be assessed at lower levels of toxicity and over a specified time period in a non-invasive manner. Biomarkers are typically measured through blood samples using bioassays to assess the biomarkers of interest. Liver toxicity is one of the most common toxicity endpoints measured since most xenobiotics are metabolised via the liver. Aspartate transaminase (AST) and alanine transaminase (ALT) are two such biomarkers associated with liver toxicity (Garg et al. 2007; Lee et al. 2012; Das et al. 2011).

2.2 Study objective

In this study we repurpose (Loshin 2010) a dataset obtained from the National Cancer Institute's Developmental Therapeutics Program (DTP) (DTP 2000). This is a dataset compiled from their in vivo screening program against various inoculated cancer cell lines to assess the effectiveness of different compounds against a developing tumour. It contains records compiled from the 1950s through to the late 1980s.

Each compound was tested by one or more laboratory (screener) at various doses against different cancerous cell lines to assess their ability as tumour reducing agents. Experiments were conducted in various strains of mice, rats and hamsters using different dosing schedules and administration routes. Test compounds were administered using a range of vehicles. As a measure of toxicity, the survival rate on a specified day was recorded.

Whilst the primary objective of the DTP was to discover anticancer compounds, we have repurposed and mined this dataset to establish drug-vehicle toxicity relationships. From this we propose a methodology to extract drug-vehicle toxicity relationships and present a means of pairwise comparisons to be made between different vehicles offering differing levels of toxicity protection. We build classifiers using C4.5 decision trees (DT) and random forests (RF) to demonstrate that our methodology can be used to build predictive models with success rates in the range of 70–80 % accuracy.

Whilst there is published work in the field of predictive toxicology, to our knowledge we are the first to build computational models based on soft computing techniques for vehicle-toxicity predictions.

3 Methodology

We propose a methodology of extracting the relevant information from our dataset and processing this using an area under the curve approach to assign classification of vehicles. This approach can be used for a multitude of problems that compare the functional relationship of two variables for different objects (vehicles in our case). Section 2.1 discusses some ways in which toxicity is commonly measured. In our study, toxicity was measured as animal survival per dose administered. Given this we are able to extract dose-survival relationships for drugs that are administered using different vehicles.

For any vehicle V^k for any given drug we consider it represented by two arrays D^k and S^k, which are of equal size, n, ($n \geq 2$) where $D^k = \left[d_1^k \ldots d_n^k\right]$ and $\mathrm{S}^k = \left[s_1^k \ldots s_n^k\right]$, D^k is the dose amount array and S^k is the survival array: $V^k = \left\{D^k, S^k\right\}$. Elements of array D^k are ordered ascendingly: $d_1^k < d_n^k$ where $n > 1$. In this case the value s_1^k represents the survival for the smallest dose amount of drug using vehicle V^k, and the value s_n^k represents the survival value for the largest dose amount of drug using vehicle V^k.

For two different vehicles (V^1 and V^2) used to test the same drug a plot of S^1 against D^1 would produce a dose versus survival curve for vehicle V^1 and likewise a plot of S^2 versus D^2 would produce a curve for vehicle V^2. Calculating their respective area under the curves can then be used to compare the toxicity difference afforded by the vehicles. Whist the elements in the arrays for V^1 may not be equivalent in length or value to the elements in the arrays for V^2 it is important to ensure that the maximum number of data points spanning the largest possible area under the dose versus survival curve is used to compare the toxicity difference between the two vehicles.

For simplicity let us consider vehicles V^1 and V^2 and the arrays they represent. To maximise data comparison these arrays were subjected to further extrapolation and interpolation before AUCs were calculated.

3.1 Extrapolation

Extrapolation was considered beyond the maximum and/or below the minimum dose for each array. For vehicles V^1 and V^2, their respective dose amount and survival values were compared. Extrapolation was considered for V^1 if its minimum dose amount was greater than the minimum dose amount of V^2 and survival at the minimum dose was 100 %. Likewise if the maximum dose amount of V^1 was less than then maximum dose amount of V^2 and survival at the maximum dose was 0 % then extrapolation was also a possibility.

Consider the set of arrays for V^1 and V^2 below:

Dose amount (mg/kg) for V^1 (D^1):

[15, 20, 30, 40, 50, 60]

% survival for V^1 (S^1):

[100, 100, 100, 70, 30, 0]

Dose amount (mg/kg) for V^2 (D^2):

[10, 20, 30, 40, 50, 60, 70]

% survival for V^2 (S^2):

[100, 100, 100, 80, 50, 30, 10]

Given that the minimum dose amount for V^1 (15 mg/kg) is greater than the minimum dose for V^2 (10 mg/kg), and the survival for V^1 at the minimum dose is 100 % then it is reasonable to assume that the survival of V^1 at a mock dose of 10 mg/kg would also be 100 %. At the opposite end of the array a similar assumption can be made. Given that the maximum dose amount for V^1 (60 mg/kg) is less than the maximum dose for V^2 (70 mg/kg), and the survival for V^1 at the maximum dose is 0 % then it is reasonable to assume that the survival of V^1 at a mock dose of 70 mg/kg would also be 0 %. The extrapolated arrays for V^1 would then look like:

(Extrapolated values underlined)

Extrapolated Dose amount (mg/kg) for V^1 (D^1):

[$\underline{10}$, 15, 20, 30, 40, 50, 60, $\underline{70}$]

Extrapolated % survival for V^1 (S^1):

[$\underline{100}$, 100, 100, 100, 70, 30, 0, $\underline{0}$]

Extrapolation could occur at both ends of the array, one end of the array, or neither end depending on the dose amounts and only if the survival values are 100 or 0 % for the minimum and maximum doses respectively.

The pre- and post- extrapolated curves for V^1 and V^2 are shown in Figs. 2 and 3 below.

Using this extrapolation method allows more data points to be used when calculating the AUCs for vehicles V^1 and V^2.

3.2 Interpolation

Interpolation was considered if extrapolation was not achievable. Namely when the minimum and maximum doses of one vehicle represented survival values of <100 or >0 % respectively. With such instances, the assumptions made during extrapolation could not be used. Therefore to maximise the

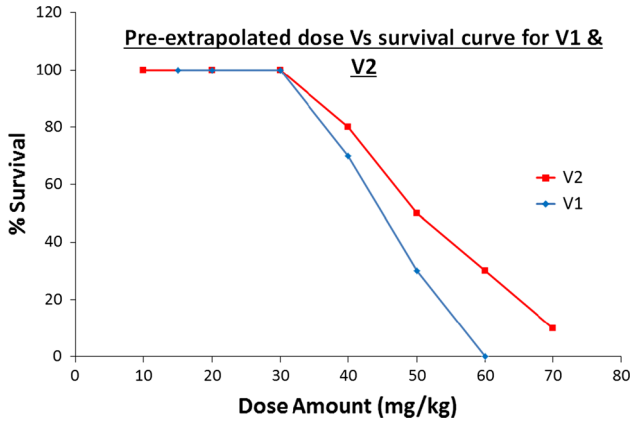

Fig. 2 Pre-extrapolated curve for V^1 and V^2

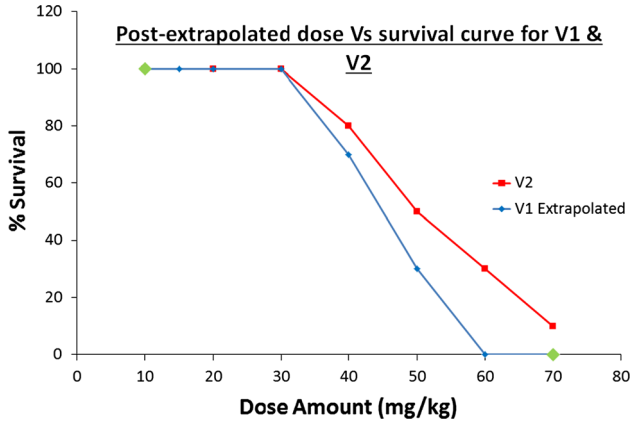

Fig. 3 Post-extrapolated curve for V^1 and V^2. The extrapolated data points for V^1 are shown in *green* (colour figure online)

number of data points we can use to calculate their respective AUCs, interpolation of data was employed.

Consider the set of arrays for V^3 and V^4 below:

Dose amount (mg/kg) for V^3 (D^3):

[5, 20, 30, 40, 50, 80]

% survival for V^3 (S^3):

[90, 80, 70, 60, 30, 10]

Dose amount (mg/kg) for V^4 (D^4):

[10, 20, 30, 40, 50, 60, 70]

% survival for V^4 (S^4):

[100, 100, 80, 70, 40, 30, 25]

The two array sets above for vehicles V^3 and V^4 are comparable over the common dose range of 20–50 mg. Since the criteria for extrapolation are not satisfied (see Sect. 3.1),

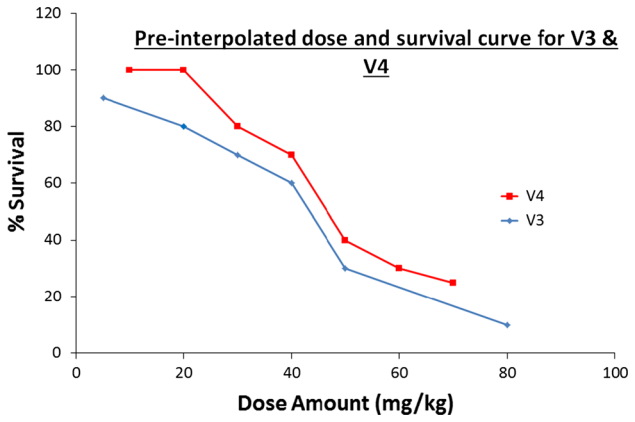

Fig. 4 Pre-interpolated curve for V^3 and V^4

linear interpolation between data points can be used. Given that the minimum dose for V^4 is 10 mg, this value can be interpolated from the data points of V^3 between the doses of 5 mg and 20 mg.

To interpolate a % survival value for V^3 (S_n^3) at a mock dose of 10 mg/kg (D_n^3) we have two dose amounts given by D_1^3 and D_2^3 and their corresponding % survival values of S_1^3 and S_2^3. The value for S_n^3 along the straight line is given by:

$$S_n^3 = S_1^3 + (S_2^3 - S_1^3)\frac{(D_n^3 - D_1^3)}{(D_2^3 - D_1^3)} \quad (1)$$

where $D_1^3 = 5$ mg/kg, $D_2^3 = 20$ mg/kg, $S_1^3 = 90$ %, $S_2^3 = 80$ % and $D_n^3 = 10$ mg/kg, then $S_n^3 = 86.6$ %

Similarly if extrapolation at the high dose end of D^3 was not possible then interpolation between two points that contained a common dose value contained within D^4 was calculated. In the example above, a value of 70 mg/kg would be interpolated into the array V^3 with its corresponding % survival value added to the array S^3.

The interpolated arrays for V^3 would then look like:

(Interpolated values underlined)

Interpolated Dose amount (mg/kg) for V^3 (D^3):

[5, 10, 20, 30, 40, 50, 70, 80]

Interpolated % survival for V^3 (S^3):

[90, 86.6, 80, 70, 60, 30, 16.6, 10]

The pre- and post- interpolated curves for V^3 and V^4 are shown in Figs. 4 and 5 below.

Both extrapolation and interpolation increase the maximum common dosage range over which two different vehicles can be compared, maximising the number of data points from the dataset that can be used.

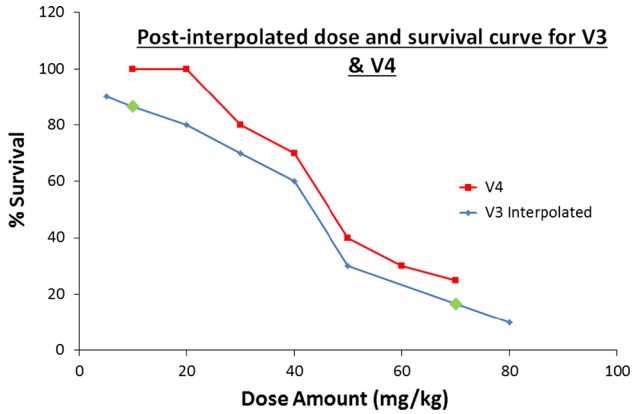

Fig. 5 Post-interpolated curve for V^3 and V^4. The interpolated data points for V^3 are shown in *green* (colour figure online)

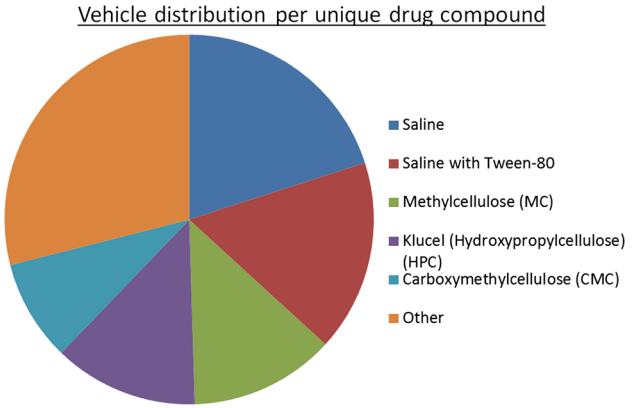

Fig. 6 Distribution frequency of the vehicles per unique drug compound within our dataset. A total of 39 vehicles were present within our dataset of which some vehicles have been used more frequently than others. The five most frequently used vehicles are shown with the remaining grouped as 'Other'

3.3 AUC calculation

After extrapolation/interpolation, the area under the dose versus % survival curve (AUC) for the two vehicles was calculated over the maximum common dosage range possible. Given the dosage range $\{D_1^n \ldots D_x^n\}$ that corresponds to a % survival array of $\{S_1^n \ldots S_x^n\}$, the AUC is calculated as follows:

$$\text{AUC} = \sum_{x=1}^{X-1} (D_{x+1}^n - D_x^n)\left(\frac{S_x^n + S_{x+1}^n}{2}\right) \tag{2}$$

AUCs for two vehicles used with the same drug can then be compared to see which provides the better toxicity protection. Over a common dose range the greater the difference in AUC between the vehicles the greater reassurance we can have that the difference is real. It is important to set a threshold for which it will be considered that the differences are significant or not. For example a difference in the AUCs of say 10 % may be taken as significant and for any comparisons that result in an AUC difference of <10 % are considered equivalent. The threshold is set with careful consideration of the data used, how reliable a researcher feels the data are and how much signal-to-noise the data may contain. For a dataset that potentially contains high noise levels, some experimentation may be necessary before a threshold can be set. For our study we set this AUC difference threshold to 30, 40 and 60 % (see Sect. 5).

4 Dataset

For this study we used the in vivo screening dataset obtained from the National Institutes of Health's Developmental Therapeutics Program (DTP) (DTP 2000). This dataset contained 2,724,199 records, detailing experimental in vivo screening results of 227,093 potential cytotoxic drug candidates. Their aim was to assess the ability of the drug candidate to reduce the size of a tumour cell line inoculated onto a host. Each drug was administered using a carrier in the form of a drug vehicle. There are 39 different vehicles used within the dataset for which some vehicles had been used more frequently than others (Fig. 6).

Whilst we do not consider our dataset to be "big data" as defined by several proposed definitions of big data, it does meet some of the criteria that are generally accepted (Laney 2001; Berman 2013; Hashem et al. 2015). Big data is not simply a large volume of data or a large structured database containing lots of records. There are several proposed definitions that currently exist to define big data of which the three V's is the most prominent.

The three V's define big data as:

Volume—this refers to the amount of the data being large and often derived from various sources.

The size of our dataset is indeed large, approximately 2.7 million records, and these data have been compiled from different sources given that there were several different laboratories (screeners) that produced these data. This gives rise to the complexity of how reliable the data from multiple sources can be, each of which will carry a different error rate associated with the data they generate making comparisons across sources challenging.

Variety—Variety means that the data come in different forms or collected via different means.

This is true to some extent with our dataset given that different laboratories (screeners) were involved in generating and reporting the data. However, the form in which these data have been compiled is the same across all screeners.

Velocity—Velocity is the speed at which the data are generated which means the dataset is dynamic due to the absorption of complementary data.

Our large dataset is not changing dynamically or increasing in size. Most of these data were generated from the early 1950s up until the late 1980s. It covers an era where computers were not widely used across industries as they are used today. A large proportion of the dataset we are working with was only available electronically due to a substantial effort made to translate all the data from a hard copy format into a digital one (DTP 2000). Our dataset does however carry the possibility of having complimentary data added to it. If similar data become available from alternative sources then there is no reason why some form of data integration cannot occur between the datasets. For our work no additional data were added to the original dataset we acquired.

Presented with some big data challenges our dataset containing approximately 2.7 million records with drug-vehicle relationships was investigated. Working with big data does not always lead to novel discoveries, but presents challenges of its own as discussed above. The main purpose of this study was to propose and develop a methodology what would enable the extraction of drug-vehicle comparisons in order to build classifiers.

5 Experimental work

5.1 Initial data curation

To make toxicity comparisons, we mined our dataset to produce dose amounts against survival plots for compounds tested using the same experimental conditions across two or more different vehicles. These vehicles were then compared pairwise. Doing this enables comparisons between the different toxicity protection levels for two different vehicles to be compared. All our data processing was accomplished using a Knime (version 2.10.1) (Knime 2000) workflow. Information held within the dataset was interpreted using a document referred to as 'Instruction_14,' available for download from the DTP website (DTP 2000). Whilst it was important to retain and keep important experimental conditions constant when making comparisons, it was also necessary to aggregate on other fields not considered to influence experimental outcome. Such fields included but were not limited to the date of the experiment or the laboratory (screener) where the experiment was conducted. Aggregating on such fields allowed a greater number of drug-vehicle comparisons to be made.

Several columns carried information about the inoculated cell line used in these experiments. We were not concerned with such information and so any columns carrying tumour cell line information were removed from the data table. Whilst it is possible that a developing tumour could affect the survival outcome of an animal, we felt that we could aggregate on such columns since survival outcome (toxday) was recorded on day 0 (zero) or day 5 after the first dose was administered and as such the toxicity outcome would be largely influenced by the drug administered rather than the developing tumour.

Table 1 List of columns retained after initial data curation

Original column heading	Description of column
NSC	Identification number of compound tested (NSC-National Service Center)
HOST_GROUP_CD	Type of animal
HOST_CD	Species/strain of animal used
ADMIN_ROUTE	Route of administration of drug being tested
INTERVAL	Time between treatment (dosing) in terms of interval unit
INTERVAL_UNIT	Minutes (M), hours (H) or days (D)
VEHICLE_CD	Vehicle used
NUMBER_INJECTIONS	Total number of injections administered
FIRST_INJECTION_DAY	Day of first administration injection relative to day zero (inoculation)
REPETITION	Repetition of injection cycle
RESTART_DAYS	Days on which repetition occurs
DOSE_AMOUNT	Dose in mg/kg unless otherwise stated
TOXDAY	Day toxicity is measured on
SURVIVOR_NUMBER_START	Animal count at start of experiment
SURVIVOR_NUMBER_TOXDAY	Animal count on toxday

The columns listed were considered influential to experimental outcome i.e. animal survival

The columns that were retained were considered relevant enough to affect the toxicity outcome of the experiments. These key columns are listed in the table below (Table 1) with their original heading name and a brief description of their value.

Upon removal of all unnecessary columns, individual records within the data table were further curated. Any records considered erroneous were removed from the data table. For example if the animal count at the end of an experiment (SURVIVOR_NUMBER_TOXDAY) exceeded the survivor count at the start of an experiment (SURVIVOR_NUMBER_START), these records were removed. Records that contained missing values in important fields were also entirely removed from the data table. Examples of this include if the vehicle (VEHICLE_CD), or dose amount (DOSE_AMOUNT) fields were empty.

Upon this initial data curation our data table was reduced to 2,710,014 records. Typically these in vivo experiments used 6 or 10 animals to begin with (SURVIVOR_NUMBER_START). To compare experiments that used 6 animals with those that used 10 we simply calculated the percentage survival rate (% survival), which was the SURVIVAL_NUMBER_END/SURVIVAL_NUMBER_START * 100. Since our dataset contained replicate experiments that resulted in slight differences in experimental outcome, it was necessary to then take a mean of the % survival values across replicate experiments (% mean survival).

To generate dose versus '% mean survival' plots we grouped data that carried the same experimental conditions, that is: NSC, HOST_GROUP_CD, HOST_CD, ADMIN_ROUTE, INTERVAL_UNIT, INTERVAL, NUMBER_INJECTIONS, FIRST_INJECTION_DAY, REPETITION, RESTART_DAYS and TOXDAY. Grouping on these fields allowed an array of lists to be constructed of the DOSE_AMOUNT, % mean survival and VEHICLE_CD fields. These lists were held as arrays that were used to plot the dose amount versus % mean survival curves. An example of the array lists generated from our dataset is shown below:

DOSE_AMOUNT (mg/kg):

[300, 1000, 3300, 10000, 300, 1000, 3300, 10000]

% mean survival:

[100, 100, 33, 0, 100, 100, 100, 67]

VEHICLE_CD:

[Saline, Saline, Saline, Saline, CMC, CMC, CMC, CMC]

CMC = Carboxymethylcellulose

Once the array lists were extracted, they were used for the pairwise comparison of two vehicles. In the example above, the array lists are split into their corresponding vehicle arrays.

DOSE_AMOUNT (mg/kg) (saline):

[300, 1000, 3300, 10000]

% mean survival (saline):

[100, 100, 33, 0]

DOSE_AMOUNT (mg/kg) (CMC):

[300, 1000, 3300, 10000]

% mean survival (CMC):

[100, 100, 100, 67]

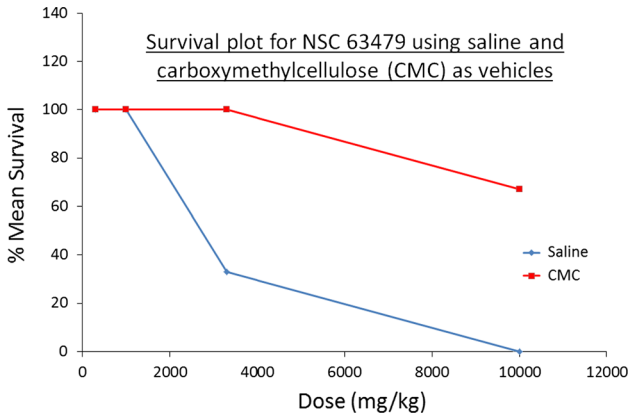

Fig. 7 The two vehicles (Saline and CMC) that show quite distinct AUC profiles when formulated with drug compound NSC 63479. At the two lower doses (300 and 1000 mg/kg) there is no observable difference between the vehicles, with both showing a survival rate of 100 %. As the doses increase to 3300 and 10,000 mg/kg it is clear that CMC provides better toxicity protection against the drug compound NSC 63479

The dose versus % mean survival curves for the saline and CMC data above is shown in Fig. 7. The Figure shows the curves generated from two different vehicles used in the formulation of drug compound NSC 63479. The curves show that for this particular compound, CMC is clearly more protective over the common dosage range of 300–10,000 mg/kg.

5.2 Classification through pairwise comparison

Using the AUC value over a common dose range for two different vehicles allowed a pairwise comparison to be made that distinguished whether one vehicle was more protective of a particular drug's toxicity than another. A difference in the AUC between two vehicles of 30 % or greater was set as a threshold to determine which vehicle was significantly better at providing toxicity relief over another. Given the challenges faced with our dataset, which was un-curated and contained noise that was difficult to measure, we set a large threshold of 30 % to ensure the differences we observed were related to vehicle differences rather than noise or experimental variation. For a particular drug that was used with the vehicles saline and CMC, if saline resulted in an AUC that was 30 % or greater than the AUC for CMC this drug was assigned to the class termed saline > CMC. Likewise if CMC resulted in an AUC of 30 % or greater, then that particular drug would be assigned to the class termed CMC > saline. For AUC differences of less than 30 %, the vehicles were considered equivalent and not assigned a class. We additionally looked at larger thresholds to assess how this may influence model performance. Thresholds of 40 and 60 % were also considered in our work (see Sect. 6).

During classification assignment there were occasions that resulted in a particular drug being assigned to more than one

class. This occurred when one set of experiments suggest that the vehicle saline is more protective than CMC, whilst for the same drug another set of experiments suggested the opposite to be true. To ensure drugs were only assigned to one class we filtered out entirely any drugs that produced such discrepancies. Removing these drugs reduced the size of our dataset but this was an import part of the final data curation before models could be built.

After all our AUC comparisons were made it was evident that although we began with 39 different vehicle types in our initial dataset, the largest number of our pairwise comparisons were between the vehicles saline and CMC. For this reason we built a binary classification model based on these two vehicles, excluding all other vehicles in our dataset for future work.

We extracted 2221 experimental records that had used both saline and CMC as a vehicle. These 2221 records corresponded to 1454 unique drug numbers (NSC). This curated dataset has been included as part of our supplementary data. Using different thresholds for AUC class assignment meant the number of drugs varied for each of the models we built using these thresholds. Of the 1454 unique drug compounds as the threshold increased from 30 to 40 to 60 % the number of compounds in the dataset decreased since more compounds would naturally be classed as 'equivalent' and so removed from the dataset. See Sect. 6, Table 2, for the number of drugs remaining in each model we built.

5.3 Chemical structure curation

To build QSAR models, chemical structure information is required for interpretation of our drugs. We obtained several chemical structure files (.sdf) from the DTP website. These were prepared at various points over the time period in which the main in vivo dataset was made electronically available. Whilst some files contained duplicate drug information which conflicted with other files, it was considered that the most recently dated file contained within it the most accurate information and so structures from this file were taken in the first instance. Any structures still not identified were then obtained from the next most recently dated sdf file and so on until as many drug structures as possible were identified. Of the 227093 unique drug numbers (NSC) in our dataset, all but 5510 were identified. Of the 1454 compounds identified which had data for both saline and CMC, 1405 had chemical structures suitable for modelling. Records for the drug structures not available were removed from the dataset.

5.4 Molecular descriptor and fingerprint generation

From the drugs successfully classified, a series of descriptors were calculated that numerically encode structural features of the compounds to be processed by our models. A total of 24 molecular property descriptors based on the Indigo tool kit (Indigo 2010) were calculated using the Molecule Properties Knime node (version 1.1.4.201308021053). Similarly MACCS fingerprints were generated using the CDK toolkit fingerprints Knime node (version 1.5.2.201409032225) which contains 166 substructure patterns denoted as binary vectors that indicate the presence or absence of a particular substructure.

5.5 Descriptor reduction

The diversity captured by these descriptors is vast but to provide a descriptor pool from which models could be built with datasets of the sizes shown in Table 2, the descriptors were objectively screened.

Descriptors which produced a constant value throughout the dataset were removed from the descriptor pool.

All but one of any two or more descriptors that were perfectly correlated were also removed. We set a mutual descriptor correlation coefficient, R^2, value of 1, but a value of 0.9 or above is reported in other studies (Eldred et al. 1999; Rodgers et al. 2010).

Lastly we correlated the binary class of our compounds with the remaining descriptor pool. We selected for the highest correlating descriptors at a 10:1 ratio of data points to descriptors. For the DTs, models were built by selecting $n/10$ descriptors when n data points existed. Whilst for the RFs, models were built by selecting $n/10$ descriptors from a pool of $2n/10$ when n data points existed.

For models built using interpolation only at the 60 % AUC threshold (see Table 2), this translates to 7 descriptors ($n/10$) given 70 data points (n) for the DTs and 7 descriptors ($n/10$) from a pool of 14 ($2n/10$) given 70 data points (n) for the RFs.

5.6 Model building

Two machine learning techniques were employed to build classification predictive models; The C.45 decision tree (DT) and random forest (RF).

C4.5 decision tree (DT) is an algorithm developed by Quinlan Shafer et al. (1996). Attributes are chosen that most effectively split the tree into their respective classes. The attribute exhibiting the highest normalised information gain is the chosen split criterion. C4.5 then repeats on the smaller sub-lists. We used no pruning for our DT models and used the Gini index as a quality measure in their building.

The random forest (RF) (Breiman 2001) is an ensemble learning method that constructs a multitude of decision trees and outputs a prediction that is the mode of the classes of the individual trees. A subset of the training dataset (local set) is chosen to grow individual trees, with the remaining samples used to estimate the goodness of fit. Trees are grown

Table 2 Model output for DT and RF using different threshold values and extrapolation processes

Model	Interpolation/ extrapolation	% AUC threshold	No. of compounds	Class split (saline/CMC)	Accuracy (%)	Correct saline (%)	Correct CMC (%)	Balanced accuracy (%)
DT	Interpolation only	30	148	97/51	60.8	70.1	43.1	56.6
DT	Extrapolation high	30	170	105/65	58.2	63.8	49.2	56.5
DT	Extrapolation high–low	30	114	67/47	64.9	67.2	61.7	64.4
RF	Interpolation only	30	148	97/51	66.2	74.2	51.0	62.6
RF	Extrapolation high	30	170	105/65	59.4	66.7	47.7	57.2
RF	Extrapolation high–low	30	114	67/47	67.5	74.6	57.4	66.0
DT	Interpolation only	40	112	70/42	64.3	77.1	42.9	60.0
DT	Extrapolation high	40	136	80/56	68.4	71.2	64.3	67.8
DT	Extrapolation high–low	40	89	49/40	73.0	73.5	72.5	73.0
RF	Interpolation only	40	112	70/42	71.4	78.6	59.5	69.0
RF	Extrapolation high	40	136	80/56	60.3	70.0	46.4	58.2
RF	Extrapolation high–low	40	89	49/40	65.2	71.4	57.5	64.5
DT	Interpolation only	60	70	45/25	71.4	84.4	48.0	66.2
DT	Extrapolation high	60	99	61/38	73.7	82.0	60.5	71.2
DT	Extrapolation high–low	60	52	31/21	65.4	67.7	61.9	64.8
RF	Interpolation only	60	70	45/25	70.0	82.2	48.0	65.1
RF	Extrapolation high	60	99	61/38	72.7	77.0	65.8	71.4
RF	Extrapolation high–low	60	52	31/21	80.8	80.6	81.0	80.8

by splitting the local set at each node according to the value of a random variable sampled independently from a subset of variables. The number of trees in our RF model was set to 100, with greater values showing no improvement.

Our dataset that contained the binary classes for Saline > CMC or CMC > Saline was then processed for vehicle prediction. We employed a 10 fold cross validation training method. This process utilises 90 % of each class to train on, whilst predicting the remaining 10 %. This was then iterated over 10 times to ensure that prediction occurs on all compounds over the 10 iterations.

A flow diagram of the entire methodology is shown in Fig. 8. The inputs are the Input Dataset and the Drug Structures which are described in Sects. 4 and 5.3. The Data Curation, Class Curation and Drug Structure Curation steps are discussed in Sects. 5.1, 5.2 and 5.3 respectively. The soft computing approaches we used for the Data Extraction and Model Training steps are described in Sects. 5.1 and 5.6. Our AUC methodology utilised Extrapolation and Interpolation techniques which are detailed in Sects. 3.1 and 3.2. Sections 3.3 and 5.4 describe how to Calculate AUC for pairwise classification and generate molecular properties and fingerprints. The Descriptor filter step shows how we screened for descriptors and is detailed in Sect. 5.5. Finally the Drug-Vehicle Models we built are described in this section (Sect. 5.6).

The Knime workflow used to build and run our models has been included in our supplementary data along with our curated dataset containing saline and CMC relationships and chemical structures.

6 Results and discussion

Using the experimental methodology discussed above we ran several prediction models. Given the importance of the classifier selection threshold (see Sect. 5.1) which in turn is determined by the interpolation and extrapolation procedures discussed in 3.1 and 3.2 of our methodology section we decided to run our models using 3 different threshold settings of 30, 40 and 60 %. For each of these 3 different thresholds we built models on data that were interpolated only with no extrapolation (Interpolation only), interpolated with high dose extrapolation (Extrapolation high) and interpolated with extrapolation at the high and low doses (Extrapolation high=-low). The output of these models is shown in Table 2. From Table 2 we see that the RF model tends to outperform the DT on almost all occasions as expected (Diaz-Uriarte and de Andres 2006; Caruana and Niculescu-Mizil 2006).

We generally see better accuracy values for DT's and RF's as the threshold increases from 30 to 40 to 60 %, possibly suggesting that the amount of noise in the original dataset is high and for any signal to be observed the necessary threshold needs to be high. We did not observe any trends in performance as the data defining the AUC range are extrapolated at one end or both—though the number of compounds included in the model is consistently greater when extrapolating at the high end only.

The best predictions we produce come from a RF model at the 60 % threshold with data extrapolated at the high and low doses (Extrapolation high–low), which produces a bal-

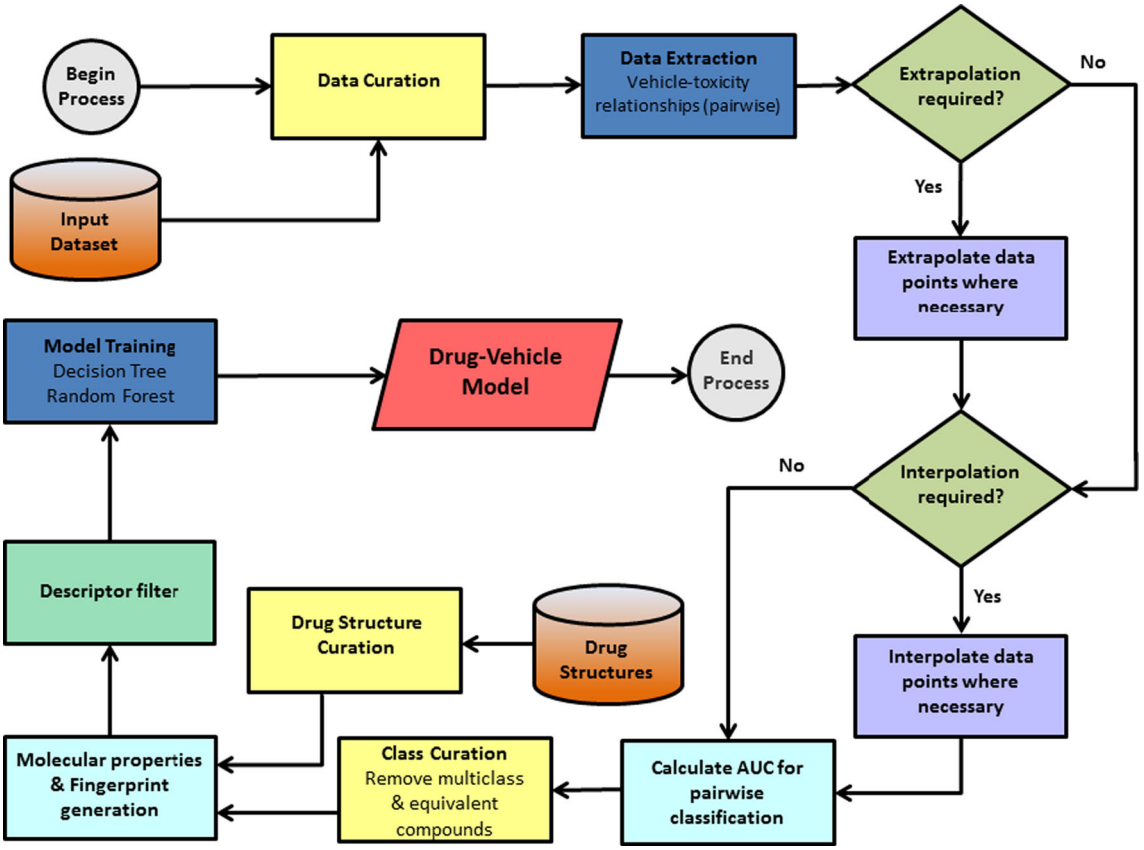

Fig. 8 Flowchart of the experimental methodology showing all processes from data curation through to model prediction. The Extrapolation and Interpolation process are iterative to completion. For simplicity, this loop is not included in the flowchart

anced accuracy of 80.8 %. This model also has individual prediction accuracies for Saline and CMC of 80.6 and 81.0 % respectively showing that the predictive outcome of the individual classes is evenly distributed. With the threshold set so high, i.e. the class discriminator for AUCs set to 60 % for classes to be assigned, we only produce a dataset of 52 compounds to be modelled.

Although the random forest models tend to outperform the decision trees we gain some insightful scientific information from the decision trees. Specifically, collecting information on how the trees are built and what descriptors the branches split on can provide some interesting scientific interpretation for our results.

Doing this for the trees built in our models brings up some interesting findings. We find that from the 9 different decision trees built in our models (each iterated 10 times) there were five MACCS fingerprints that were used frequently to split the branches on. These five most common fingerprints were (position-description); 120-HETEROCYCLIC ATOM >1, 109-ACH2O, 115-CH3ACH2A, 137-HETEROCYCLE and 53-QHAAAQH.

7 Conclusion

In this study we have introduced a methodology that shows how classifiers can be built on scientific data for model predictions. We explain the steps of this methodology that is applicable across many scientific disciplines. We show how existing scientific data can be repurposed and mined for new knowledge discovery. An additional innovative feature of our methodology is the use or our proposed interpolation and extrapolation to increase the data points used for real world datasets.

For our experimental work we acquired a large dataset from the Developmental Therapeutics Program and showed our methodology to work well and built classifiers that produce accuracies of 80 %.

This work shows that the effects on toxicity of drugs by different vehicles can be modelled, and that well-performing, interpretable models can be built if sufficient data are available. To the best of our knowledge this is the first study building models in this field.

We consider for our future work, extending our current approaches in dataset preparation to non-linear interpolation

and extrapolation to deal with survival curve complexity. We also intend to explore other vehicle relationships that are contained within our dataset.

Acknowledgments We are grateful to Lhasa Limited for a grant for PhD funding for Pritesh Mistry. We would also like to thank Daniel Zaharevitz of NIH for helping to make the data readily accessible.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ajith TA et al (2007) Ascorbic acid and alpha-tocopherol protect anticancer drug cisplatin induced nephrotoxicity in mice: a comparative study. Clin Chim Acta 375(1–2):82–86

Arima H et al (2001) Comparative studies of the enhancing effects of cyclodextrins on the solubility and oral bioavailability of tacrolimus in rats. J Pharm Sci 90(6):690–701

Bakhtyari NG et al (2013) Comparison of in silico models for prediction of mutagenicity. J Environ Sci Health Part C Environm Carcinog Ecotoxicol Rev 31(1):45–66

Basavaraj S, Betageri GV (2014) Can formulation and drug delivery reduce attrition during drug discovery and development–review of feasibility benefits and challenges. Acta Pharm Sin B 4(1):3–17

Berman JJ (2013) Principles of big data

Blix HS et al (2010) Drugs with narrow therapeutic index as indicators in the risk management of hospitalised patients. Pharm Pract 8(1):50–55

Breiman L (2001) Random Forests. Mach Learn 45(1):5–32

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning, ACM, pp 161–168

Cruz-Monteagudo M et al (2008) Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. J Comput Chem 29(4):533–549

Das S et al (2011) Silymarin nanoparticle prevents paracetamol-induced hepatotoxicity. Int J Nanomed 6:1291–1301

Developmental Therapeutics Program (DTP) (2000). http://dtp.nci.nih.gov/

Diaz-Uriarte R, de Andres SA (2006) Gene selection and classification of microarray data using random forest. Bmc Bioinformatics 7

Eldred DV et al (1999) Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. Chem Res Toxicol 12(7):670–678

Eldred DV, Jurs PC (1999) Prediction of acute mammalian toxicity of organophosphorus pesticide compounds from molecular structure. Sar Qsar Environ Res 10(2–3):75–99

Garg M et al (2007) Reduced hepatic toxicity, enhanced cellular uptake and altered pharmacokinetics of stavudine loaded galactosylated liposomes. Eur J Pharm Biopharm 67(1):76–85

Hans ML, Lowman AM (2002) Biodegradable nanoparticles for drug delivery and targeting. Curr Opin Solid State Mater Sci 6(4):319–327

Hashem IAT et al (2015) The rise of "big data" on cloud computing: review and open research issues. Inf Syst 47:98–115

Hodgman MJ, Garrard AR (2012) A review of acetaminophen poisoning. Crit Care Clin 28(4):499–516

Huo M et al (2010) Synthesis and characterization of low-toxic amphiphilic chitosan derivatives and their application as micelle carrier for antitumor drug. Int J Pharm 394(1–2):162–173

Hutchinson L, Kirk R (2011) High drug attrition rates-where are we going wrong? Nat Rev Clin Oncol 8(4):189–190

Indigo knime node tool kit (2010). http://sourceforge.net/projects/cdk/

Injac R et al (2008) Cardioprotective effects of fullerenol C-60(Oh)(24) on a single dose doxorubicin-induced cardiotoxicity in rats with malignant neoplasm. Technol Cancer Res Treat 7(1):15–25

Italia JL et al (2007) PLGA nanoparticles for oral delivery of cyclosporine: Nephrotoxicity and pharmacokinetic studies in comparison to Sandimmune Neoral (R). J Control Release 119(2):197–206

James LP et al (2003) Acetaminophen-Induced hepatotoxicity. Drug Metab Dispos 31(12):1499–1506

Kaminskas LM et al (2012) Doxorubicin-conjugated PEGylated dendrimers show similar tumoricidal activity but lower systemic toxicity when compared to pegylated liposome and solution formulations in mouse and rat tumor models. Mol Pharm 9(3):422–432

Kelava T et al (2010) Influence of small doses of various drug vehicles on acetaminophen-induced liver injury. Can J Physiol Pharmacol 88(10):960–967

Kim SC et al (2001) In vivo evaluation of polymeric micellar paclitaxel formulation: toxicity and efficacy. J Control Release 72(1–3):191–202

Knime (2000). http://www.knime.org/

Laney D (2001) 3D data management: controlling data volume, velocity, and variety. Meta Group

Larabi M et al (2004) Study of the toxicity of a new lipid complex formulation of amphotericin B. J Antimicrob Chemother 53(1):81–88

Lee NH et al. (2012) Hepatoprotective and antioxidative activities of cornus officinalis against acetaminophen-induced hepatotoxicity in mice. Evidenc Based Complem Altern Med

Lee SST et al (1996) Role of CYP2E1 in the hepatotoxicity of acetaminophen. J Biol Chem 271(20):12063–12067

Lee HC et al (1999) Protective effect of chlormethiazole, a sedative, against acetaminophen-induced liver injury in mice. Korean J Intern Med 14(2):27–33

Lee ES et al (2003) Polymeric micelle for tumor pH and folate-mediated targeting. J Control Release 91(1–2):103–113

Liang BA et al (2013) Illegal "No Prescription" internet access to narrow therapeutic index drugs. ClinTher 35(5):694–700

Liu Y et al (2011) Dual targeting folate-conjugated hyaluronic acid polymeric micelles for paclitaxel delivery. Int J Pharm 421(1):160–169

Loftsson T (1998) Increasing the cyclodextrin complexation of drugs and drug bioavailability through addition of water-soluble polymers. Pharmazie 53(11):733–740

Loshin D (2010) The Practitioner's Guide to Data Quality Improvement, Morgan Kaufmann

Low Y et al (2011) Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. Chem Res Toxicol 24(8):1251–1262

Maliakel DM et al (2008) Prevention of cisplatin-induced nephrotoxicity by glucosides of ascorbic acid and alpha-tocopherol. Experim Toxicol Pathol 60(6):521–527

Modi S et al (2012) Integrated in silico approaches for the prediction of Ames test mutagenicity. J Comput Aided Mol Des 26(9):1017–1033

Mora LD et al (2003) The effects of oral glutamine on cisplatin-induced nephrotoxicity in rats. Pharmacol Res 47(6):517–522

Patten CJ et al (1993) Cytochrome-P450 enzymes involved in acetaminophen activation by rat and human liver-microsomes and their kinetics. Chem Res Toxicol 6(4):511–518

Pereverzeva E et al (2007) Influence of the formulation on the tolerance profile of nanoparticle-bound doxorubicin in healthy rats: Focus on cardio- and testicular toxicity. Int J Pharm 337(1–2): 346–356

Porter CJH et al (2007) Lipids and lipid-based formulations: optimizing the oral delivery of lipophilic drugs. Nat Rev Drug Discov 6(3):231–248

Rodgers AD et al (2010) Modeling liver-related adverse effects of drugs using kNearest neighbor quantitative structure activity relationship method. Chem Res Toxicol 23(4):724–732

Savjani KT et al (2012) Drug solubility: importance and enhancement techniques. ISRN Pharm 2012:195727–195727

Shafer J et al. (1996) SPRINT: a scalable parallel classifier for data mining. In: Proceedings of the international conference on very large data bases, pp 544–555

Shin H-C et al (2009) Multi-drug loaded polymeric micelles for simultaneous delivery of poorly soluble anticancer drugs. J Control Release 140(3):294–300

Uchino H et al (2005) Cisplatin-incorporating polymeric micelles (NC-6004) can reduce nephrotoxicity and neurotoxicity of cisplatin in rats. Br J Cancer 93(6):678–687

Xu C et al (2012) In silico prediction of chemical ames mutagenicity. J Chem Inf Model 52(11):2840–2847

Yu BG et al (1998) Polymeric micelles for drug delivery: solubilization and haemolytic activity of amphotericin B. J Control Release 53(1–3):131–136